

## Modelling Clustered Data in Autobiographical Memory Research: The Multilevel Approach

DANIEL B. WRIGHT\*

*University of Bristol, UK*

### SUMMARY

Much memory research involves recording several autobiographical memories for each of several people. These memories are not independent of each other, an assumption of the statistical procedures used in many cognitive psychology papers. In recent years there have been both statistical and computational advances for modelling these hierarchical data structures. This is often called *multilevel modelling*. Using data from recent memory research (Burt *et al.*, 1995), I describe this approach and show how it compares favourably with traditional approaches. © 1998 John Wiley & Sons, Ltd.

*Appl. Cognit. Psychol.* **12**: 339–357 (1998)

To explore autobiographical memory, researchers often ask people to remember several different events. These events will, in some way, be sampled from each person's life. The events are usually different for each person. Conceptually, these events are *nested* within each person. By this, I mean, that each event happens only for a single person. The data structure is called hierarchical or multilevel. There are many considerations for analysing data of this form. The purpose of this paper is to describe these considerations and to compare alternative approaches. Data from a recent study by Burt and colleagues (1995) are used to illustrate these approaches.

There are four main approaches that someone may take to analyse data of this type. The first approach ignores that events are clustered within each person and treats all event memories as statistically independent. The second approach involves calculating aggregate measures for each person, like the means over a number of events, and using these values in the analyses. This is sometimes called the 'means as outcomes' approach. For the third approach, dummy variables are made for each person. In its simplest form this is much like an analysis of covariance, where the subjects are treated as covariates. The final approach is the *multilevel approach*, in which the events are treated as nested within subjects, and both subjects and memories are treated as random variables.

I concentrate on the multilevel modelling approach because this approach will be new to many readers. Multilevel modelling, under several names, is becoming

\*Correspondence to: Daniel B. Wright, Department of Experimental Psychology, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK. E-mail: D.B.Wright@Bristol.ac.uk

Contract grant sponsor: British Academy Fellowship

increasingly popular for analysing clustered data in education, sociology, geography, and the medical sciences. It has also been used in developmental (Bryk and Raudenbush, 1987), organizational (Vancouver *et al.*, 1994) and forensic psychology (Wright and McDaid, 1996), and has great potential for memory research. Trying to analyse data in this form with traditional methods is difficult for various reasons described throughout this paper. Multilevel modelling solves some of the problems, although other concerns are raised. Several books on multilevel modelling exist (for example, Bryk and Raudenbush, 1992; Goldstein, 1995; Hox, 1995; Kreft and de Leeuw, 1998; Longford, 1993; Woodhouse *et al.*, 1995).

It is worth introducing the concept of multilevel modelling through an example. Suppose you were examining the impact of bullying on exam scores of children. You would probably first sample classrooms and then pupils within these classrooms. If you had 100 classrooms with ten pupils questioned within each, the first question that a researcher might ask would be what is the unit of analysis: pupils or classrooms.

The first approach would treat the pupils as  $n = 1000$  independent units. However, with bullying, the probability of one child being bullied is higher if other children in the class are bullied. Further, the effects of bullying are likely to influence classroom behaviour and teaching quality for the entire class, not just of the direct victim(s) of the bullying. This means that the data are not independent. While the effects of clustered data on standard errors are well known (see Scariano and Davenport, 1987), and various design effect adjustments can be made, this does not get around the problem that there are interesting substantive effects at both the pupil and classroom levels.

Given this information, many researchers might opt for the second approach, which treats classrooms as the appropriate units ( $n = 100$ ). When the units are classrooms, the hypotheses that can be explored are necessarily different from those that can be explored with pupils as the units. A hypothesis that might be considered is that classrooms with lots of problems with bullying also score low, as a group, on academic performance. The decrease in sample size will worry many researchers, but of more conceptual difficulty is that the hypotheses have changed and no longer are the effects on individuals being explored. It is possible that the hypotheses of interest will be about classroom effects, but researchers must be aware that these are the effects they are modelling.

In general, inferring characteristics at one level from analyses at another is often misleading. Making inferences about groups from individual relationships is called the *atomistic fallacy*. Making inferences about individual relations from grouped data is called the *ecological fallacy*, or sometimes the *Robinson effect* (Robinson, 1950; see Perfect, 1994, for related discussion in gerontology). The ecological fallacy is the more common error. Using data from the 1930s, Robinson compared the percentage of blacks in nine US geographical areas to the literacy levels in these areas. The correlation for these nine areas was 0.95, while the correlation at the individual level was about 0.20. A topical example of the ecological fallacy for academics is provided by Kreft *et al.* (1995; see also Kreft and de Leeuw, 1998). They discuss how, at the individual level, there is a high association between level of education attainment and salary, but that the association is much lower between places with many people with high attainment and mean salaries of these places. This is because schools and universities have people with lots of degrees but the employees are not paid as well as many other professions.

The correlations can even be in the opposite direction. In psychology, we often think of responses to particular questions as nested within an individual. For individuals, there is an accuracy and response-time trade-off (i.e. a negative within-individual correlation), but on many tasks, people who are accurate are also fast simply because they have expertise at the task (i.e. a positive between-individual correlation). The important point for this paper is that the relationships being examined are different and researchers should not infer aspects of one from estimates of the other.

The third approach is to treat classrooms as a fixed effect. For this example it would involve creating 99 dummy variables, and putting them into the model, along with the intercept, to stand for the one hundred classrooms. This has the problem that an additional 99 parameters are estimated, making it a less parsimonious model. Because of this, the estimates are more dependent on sampling variation and therefore less reliable. Increasing the number of parameters can also cause problems for estimation. Conceptually, there is the problem that because the classrooms are no longer being treated as a random sample from some population, inference to other classrooms should be tentative. Further, this approach is limited because it assumes that the slopes of the relationship between bullying and academic scores are the same for each classroom (i.e. the classrooms' regression lines are parallel). This constraint can be relaxed by having an additional 99 parameters estimated for the different slopes, but this increases the parsimony problems. This lack of parsimony is because nothing is being hypothesized about the relationships among the intercepts (or slopes) for the different classrooms.

It will be useful to introduce some notation at this point that will be used throughout this paper. Suppose  $Y_{ij}$  is a measure of academic performance for the  $i$ th pupil in the  $j$ th classroom,  $X_{ij}$  is some measure of being victimized by bullying for each pupil,  $\beta_0$  is the intercept and  $\beta_1$  is the slope. The regression for the first approach is  $Y_{ij} = \beta_0 + \beta_1 X_{ij} + e_{ij}$ , where we assume that the  $e_{ij}$  are independent and usually that they are Normally distributed. The second approach involves finding the mean of  $Y_{ij}$  and  $X_{ij}$  for each of the  $j$  classrooms, call these  $Y_j$  and  $X_j$ , and regressing the means, such that  $Y_j = \beta_0 + \beta_1 X_j + u_j$ , where  $u_j$  is the residual for each classroom. The fixed intercept approach involves creating 100 different  $\beta_0$  values, one for each classroom. I will denote fixed effects with superscripts in parentheses, thus the  $J = 100$  intercepts would be  $\beta_0^{(j)}$ . Its equation is  $Y_{ij} = \beta_0^{(j)} + \beta_1 X_{ij} + e_{ij}$ , and also assumes that the  $e_{ij}$  are independent and Normally distributed.

With the fourth approach, multilevel modelling, information about these intercepts is hypothesized. Let  $\beta_{0j}$  be a random variable and make some assumptions about its distribution. For multilevel linear regression, it is usually assumed that the intercepts for the different classrooms are centred on some  $\beta_0$  with residual  $u_j$ , that they are Normally distributed and that they are independent (and if not, another level in the hierarchy can be added). The equation becomes  $Y_{ij} = (\beta_0 + u_j) + \beta_1 X_{ij} + e_{ij}$ , or separating the fixed and random parts of the model,  $Y_{ij} = (\beta_0 + \beta_1 X_{ij}) + (u_j + e_{ij})$ . More complex models, like letting  $\beta_1$  vary for the different classrooms, can be incorporated. An important difference between this approach and the third approach is the number of parameters that need estimating. The third approach requires estimating 100 parameters for  $\beta_0^{(j)}$  while only two parameters are needed for estimating  $\beta_0 + u_j$  (one for the mean and one for the variance, which assuming Normality completely describes the distribution). This makes multilevel models much

more parsimonious than fixed-effect models. Also, it is easier to justify making generalizations to other classrooms since they are treated as a random sample from some population of classrooms.

Some of the popular statistical packages traditionally have had trouble modelling regressions of this form (i.e., with two or more random variables).<sup>1</sup> This, rather than any conceptual complexities, has restricted the popularity of these models in psychology. Several efficient ways of estimating these equations have been devised in recent years (see Goldstein, 1995, sections 2.5 and 2.6, for discussion). This paper is designed to introduce and to encourage memory researchers who collect multilevel data to consider using these procedures. The aim is not to provide details of the algorithms.

Using multilevel modelling could have additional value because it may help researchers to look at their theories in different ways. Gigerenzer (1991) has shown that much theorizing in social and cognitive psychology has used statistical techniques as metaphor for theory. For example, multilevel modelling may help to differentiate mechanisms that operate at the individual level and those that operate on event representations within the individual. In this paper I described a study by Burt and colleagues (1995). I go through the four approaches outlined above and compare their value.

### SNAPSHOTS OF MEMORIES

One difficulty faced by people who research long-term autobiographical memory is sampling events, or memories, from people's lives which are in some way representative of some larger population of events or memories. Some good examples of researchers sampling events/memories include having subjects keep diaries (Conway *et al.*, 1996; Larsen, 1992; Thompson, 1982; Thompson *et al.*, 1993; Wagenaar, 1986), finding subjects who have kept diaries for years (Burt, 1992), having subjects record what they are doing when a beeper goes off (Brewer, 1988), asking subjects to recall memories after giving them non-specific prompts (Crovitiz and Schiffman, 1974; Robinson, 1976), using people's own memory descriptions to prompt other memories (Brown, this issue) and asking subjects to recall their most vivid memories (Rubin and Kozin, 1984). When studying autobiographical memory, the events and the memories are unique for the person and this is important for theories of autobiographical memory (Conway, 1993). The events are *nested* within the person in a manner analogous to how pupils are *nested* within classrooms in the earlier example.

To illustrate the statistical approaches I use an excellent example of autobiographical memory research by Burt and colleagues (1995). At the beginning of the summer holiday, 27 volunteers were given several rolls of film with the instructions that they should take photographs as they normally would. Several thousand photographs were taken. Some were excluded because they were out of focus or one of several depicting the same event (for example, many pictures of a wedding ceremony). I have also excluded some for missing values. The number of photographs kept for analyses is 1342. The number per person ranged from 23 to 147. Photographs are

<sup>1</sup>SAS PROC MIXED is able to handle some of the problems discussed here. Other mainstream statistical packages are increasing their capabilities for having multiple random variables. For example, Raudenbush (1995) has shown how for some multilevel problems, structural equation software can be used to estimate the solutions.

nested within people. The photographs were coded by whether they were of an activity, of participants, of a location, or some combination of these.

After the summer holiday, these photographs were presented to subjects for 10 seconds on a tachistoscope. Several distractors, taken from the researchers' own collections, were also presented. I will not discuss these. Over 97% of them were correctly recognized as foils. Subjects were asked if they could remember the event and the reaction times were recorded. If they could remember the event, they made several ratings about it, including the importance of the event and the level of emotion provoked by the event at the time. If they could not remember the event, they were asked whether they thought it was a distractor, whether there were too many similar events to be sure, or whether the picture did not provide sufficient cues for them to decide if it was their photograph.

Burt *et al.* (1995) explored several aspects of these data. Here I examine only three relationships. While these were all important for their study, I have chosen them to cover three of the main statistical procedures used by cognitive psychologists. The first is comparing reaction times by whether the event was correctly recognized, or not, and if not recognized the reason they gave. Burt *et al.* used an ANOVA for this. The second is the relationship between importance and emotion. Both variables were measured on 7-point rating scales and the Pearson correlation was found in the original analyses. The final relationship is between the type of photograph (of activities, participants and/or locations) and whether the subject could remember the event. Burt *et al.* reported percentages for these with no inferential modelling.

### Reaction times

There was a missing value for one reaction time, making the  $n = 1341$  for these analyses. The reaction times were highly skewed (3.55,  $se = 0.07$ ). There were some that were longer than 10 seconds, which was the duration that the picture was presented on the tachistoscope. Burt *et al.* (1995) excluded these. I included them but have recoded these times as 10 seconds. The natural logarithms of these data were taken. These transformations removed most of the skewness (0.35,  $se = 0.07$ ) and made the distribution appear roughly Normal (this is different from what Burt *et al.* did, therefore the *ns* and estimates will be different here than in their analyses). I call this variable *Intime*.

About 65% of the holiday photographs were remembered ( $n = 879$ ). This will serve as the baseline category with dummy variables for distractors ('dist'  $n = 40$ ), those not recalled because there were others that were too similar ('simil'  $n = 222$ ) and those for which the cues were insufficient ('cues'  $n = 200$ ). Besides 'remembered' being the most frequent category, and therefore providing more reliable parameter estimates, the most logical contrasts are between 'remembered' and each of the others.

The first approach, which is equivalent to an ANOVA (Cohen, 1968), is a linear regression with these three dummy variables:

$$\text{Intime}_i = 0.44 + 0.32 \text{ cues}_i + 0.30 \text{ dist}_i + 0.22 \text{ simil}_i + e_i$$

(0.02) (0.05) (0.10) (0.05)

The parameter estimates and their standard errors (in parentheses below the estimates) show that the reaction times are faster for the remembered events. The 95%

confidence intervals of the parameters for the three dummy variables do not overlap with zero. Overall, the model fits significantly better than chance ( $F(3,1337) = 20.58$ ;  $p < 0.001$ ), but  $R^2$  is only 0.04 and therefore the differences cannot be considered large.

The second approach involves treating the subjects as the units under investigation. This can be done in several ways, but I will consider the two most common ways. The first is by calculating the mean log reaction time of each individual for each of the four responses and comparing these. There are questions about what to do with people who did not make one of the responses and therefore have no mean time for this response type.<sup>2</sup> Ten people did not say any of their pictures were distractors, and a couple had no photographs that they could not recognize because of too many similarities or lack of cues. When dealing with almost any data set, decisions must be made about how to deal with missing values. If subjects with any missing data are excluded from the analysis (listwise deletion), this would reduce the size of the sample of subjects and therefore lower the power (and may present bias if these excluded people differ systematically from the others). If the mean for the subject was given (and this might be a sensible technique here because of the amount of between subject variation), this also decreases the power. If the mean of the condition is given, this increases the chance of a Type I error. There is a great deal of discussion about dealing with non-response (see Little, 1992; Rubin, 1987). Here, I have taken a conservative approach and excluded missing data listwise (new  $n = 15$ ).

This analysis also showed that the means differed ( $F(3,42) = 4.36$ ,  $p = 0.009$ ), but the model accounted for 24% of the variation. This sixfold increase demonstrates the ecological fallacy and shows that aggregate data should not be used to estimate effects at lower levels. The 'remember' responses were the fastest (mean = 0.53, se = 0.08), but differences were only significant between them and the 'not enough cues' (mean = 1.00, se = 0.18) and 'distractor' (mean = 0.99, se = 0.20) responses, not the 'too similar' responses (mean = 0.72, se = 0.16).

Another way to aggregate the data is by calculating the proportion of three of the memory responses (the fourth is unnecessary since it is one minus the others) for each person and using these proportions to predict the overall mean log reaction time for each subject. This may seem a very roundabout way of assessing an effect. For many research questions it is, but I describe it because similar procedures are often used. Here it is the same as saying that people who tend to make certain responses to the memory question also tend to have different reaction times. The result for this comparison was non-significant ( $R^2 = 0.06$ ,  $F(3,23) = 0.46$ ,  $p = 0.71$ ). As Hand (1994) has elegantly described, it is critical to consider whether the statistical model being evaluated is the same as the theoretical one of interest. Here it is not.

The third approach involves treating each subject as a fixed effect. Twenty-six dummy variables were created for the subjects. When the other variables, 'cues', 'dist' and 'simil', are included in the model, the subject dummy variables (and the intercept) will stand for the mean reaction times for the remembered events for each subject (several other codings could be used, but these will suffice). First, just the 26 subject dummy variables and the constant (which will denote the 27th subject) were entered and the result was  $R^2 = 0.42$  ( $F(26,1314) = 36.62$ ;  $p < 0.001$ ). When the additional

<sup>2</sup>There is also the related problem that a mean of a category with a single response is treated as precise, and given equal weight to, a mean of ten responses.

three variables were added,  $R^2$  rose to 0.45 ( $F(29,1311) = 36.27$ ,  $P < 0.001$ ; change  $F(3,1311) = 19.72$ ,  $p < 0.001$ ). While statistically significant, the additional proportion explained beyond the subject dummy variables is not large either in absolute terms or with respect to the proportion explained by the differences among the subjects. The equation, where  $\beta 0^{(J)}$  stands for the 26 dummy variables and the intercept (i.e.  $J = 27$ ) and  $e_i$  are the residuals, is (standard errors shown below estimates):

$$\text{Intime}_i = \beta 0^{(J)} + 0.27 \text{ cues}_i + 0.29 \text{ dist}_i + 0.16 \text{ simil}_i + e_i$$

(0.04)            (0.08)            (0.04)

Additional dummy variables could be created to allow for differential effects for each subject, though this would mean about 100 parameters would need to be estimated.

The final approach, multilevel modelling, also allows the intercept, here the times for remembered photographs, to vary among subjects but in addition it assumes that these response times are Normally distributed. Rather than solve for each subject, this approach only estimates the mean intercept ( $\beta 0$ ) and its variance ( $\text{var}(u_j)$ ), because these two parameters completely describe the distribution. For all the multilevel modelling discussed in this paper, the package MLn is used (Woodhouse, 1995). Other packages designed for this type of data exist and some of these are reviewed in Kreft *et al.* (1994; see also Goldstein, 1995, chapter 11). Here the model is

$$\text{Intime}_{ij} = \beta 0 + \beta 1 \text{ cues}_{ij} + \beta 2 \text{ dist}_{ij} + \beta 3 \text{ simil}_{ij} + u_j + e_{ij}$$

where  $\beta 0$  is the mean intercept and  $u_j$  is the variation around it for the 27 subjects. When this is estimated in MLn (restrictive iterative generalized least squares, RIGLS, is used throughout this paper (see Goldstein, 1989, for technical details)), the result is

$$\text{Intime}_{ij} = 0.45 + 0.27 \text{ cues}_{ij} + 0.29 \text{ dist}_{ij} + 0.17 \text{ simil}_{ij} + u_j + e_{ij}$$

(0.08) (0.04)            (0.08)            (0.04)

with  $\text{var}(u_j) = 0.18$  ( $se = 0.05$ ) and  $\text{var}(e_{ij}) = 0.23$  ( $se = 0.01$ ). The fixed estimates are similar to the first approach. The variance of the subject level residuals ( $u_j$ ) is more than three times its standard error. The size of  $\text{var}(u_j)$  shows that it is important to take into account the variation among subjects; the assumption of independence of the first approach is not valid. In MLn the fit of a model is assessed by a  $\chi^2$  statistic. Here is it 1916.80 and this value can be compared with other models.

Another model that is worth exploring is where the variances are allowed to differ for the different memory responses. In much research, the variance of reaction times increases as the reaction times increase (sometimes logging the data prevents this, sometimes not). Because the different response categories are associated with different reaction times, it is worth including this in the model. It is predicted that the variance will be higher for the non-remembered photographs than for the remembered photographs. To investigate this a dummy variable was created with 0 for the remembered photographs and 1 for all others. The variable is called 'not\_mem'. Two random variables,  $e1_{ij}$  and  $e2_{ij}$ , are used for the photographs. The random part of the model is ( $u_j + e1_{ij} + \text{not\_mem}_{ij} * e2_{ij}$ ) so that  $\text{var}(e2_{ij})$  will be the difference in variance between remembered and not-remembered photographs.

The estimates of the fixed parameters are essentially unchanged. The new variance estimates (standard errors in parentheses) for the three random variables are  $\text{var}(u_j) = 0.17$  (0.05),  $\text{var}(e1_{ij}) = 0.18$  (0.01) and  $\text{var}(e2_{ij}) = 0.07$  (0.01). The last of these refers to the difference in variation between remembered and not-remembered photographs. The fit of this model is  $\chi^2 = 1870.50$ , a difference from the last model of  $\chi^2(1) = 46.30$ ,  $p < 0.001$ . Therefore, this term for heteroscedasticity should be included in the model.

Further extensions to this basic multilevel ANOVA model can be easily made. These include factorial designs, within-subject designs, heteroscedasticity at the subject level, further hierarchies (like subjects nested within interviewers), and many others.

### Importance and emotion

In much memory research the importance of an event and people's emotional reactions to the event are associated (e.g. Rubin and Kozin, 1984; Wright *et al.*, 1998). It is not particularly surprising that these are positively correlated. Most research that examines this relationship simply points out that they are correlated for the events used, but cannot go further to estimate the size of this relationship for some greater population of events because the events used were not chosen to be representative of any population. Burt *et al.*'s (1995) study is valuable because it is possible to argue that the events are representative of some larger population. Therefore the importance/emotion relationship can be explored in a more meaningful way than simply stating that it is unlikely to be absent (the null hypothesis significance testing approach).

Subjects only made ratings for the events they remembered. Of the 879 remembered photographs, there was one missing value for emotion reducing the number of photographs to 878. Both ratings were on 7-point scales. Treating these as continuous interval measures makes some assumptions. Alternatives, for example, re-scaling the data (Van der Geer, 1993a,b), treating them as ordered categories within a log-linear (or log-multiplicative) framework (Agresti, 1989; Clogg and Shihadeh, 1994) or examining strictly ordinal hypotheses (Cliff, 1993, 1994) are all possible. Each of these constitutes another approach which has been recently reviewed. I will use the unscaled data to keep with common practice in psychology research.

Approach one is a simple linear regression between these two ratings. While arguments could be made either for importance or for emotion to be affected by the other, I have arbitrarily chosen emotion to be the response variable and importance to be the predictor variable. The mean rating for importance was 3.67. It is helpful during the modelling to centre this variable so that the intercept is more meaningful (see Krefth *et al.*, 1995, for further issues about centering). This was done for all analyses in this section and in the Appendix, though I untransform the variable when graphing the data. Centering was not necessary for the first example since the predictor variables were all dummy variables and thus the estimates could be easily interpreted. The result of the regression is  $R^2 = 0.40$  ( $F(1,876) = 589.79$ ,  $p < 0.001$ ),<sup>3</sup> with the estimated equation of:

$$\text{emot}_i = 4.54 + 0.57 \text{ import}_i + e_i$$

(0.04) (0.02)

<sup>3</sup>This is a different value from that found by Burt *et al.* (1995) because they report only the correlation for photographs that depicted a location, an activity and people.

Because importance has been centered,  $\beta_0$  is mean value of emotion (this equality is true only for this regression). Compared with the reaction time data, this model accounts for a substantial proportion of the variation.

The second approach treats subjects as the unit of analysis. The hypothesis would be that people who on average give high ratings of importance also give high ratings of emotion. This situation is the most common for amalgamated analyses and it provides probably the most substantively interesting of the amalgamated hypotheses explored in this paper. However, it still has some drawbacks and should not be used to infer relationships about the actual events. While this hypothesis *could* reflect some causal connections between importance and emotion for events, it may simply reflect that people who have particularly important summers also have emotional summers. It is also possible that this could reflect a methodological artefact. Saris (1988) has described mathematical models for analysing differences in the response functions people have for mapping their 'true' beliefs onto response scales (see Wright *et al.*, 1994, for an example of this). Certain people may simply prefer using higher or lower regions of both rating scales for the same true levels of importance and emotion. Even without there being a relationship between importance and emotion, this would produce a high positive correlation. Therefore, the results of these aggregate data could be measuring some prolonged period of importance/emotion, an artefact and/or a relationship between an event's importance and emotion. The parameter estimates are similar to those produced by the first approach:

$$\text{emot}_j = 4.47 + 0.59 \text{ import}_j + u_j \\ (0.12) \quad (0.09)$$

The proportion of variation accounted for ( $R^2 = 0.61$ ;  $F(1,25) = 38.38$ ,  $p < 0.001$ ) is much larger than found with the first approach.

The third approach involves entering the 26 subject dummy variables into the equation while treating each photograph as a unit of analysis. This produces  $R^2 = 0.41$  ( $F(26,851) = 22.41$ ,  $p < 0.001$ ). When the variable for importance is added, the result is  $R^2 = 0.59$  which is a substantial and highly significant increase ( $p < 0.001$ ). The resulting equation is

$$\text{emot}_j = \beta_0^{(j)} + 0.60 \text{ import}_j + e_j \\ (0.03)$$

which is similar to the results from the first approach.

The above equation allows each subject to have a different intercept for emotion. Suppose we let each subject also have their own slope (i.e. relax the parallel line constraint). While this increases the complexity of the model by a further 26 degrees of freedom, it has the conceptual advantage that all that is left for the residuals, the  $e_i$ , is within-subject variation (at a total cost of estimating 54 parameters). Conceptually, this is similar to running an ordinary least squares regression for each subject. If we do this we get  $R^2 = 0.64$  ( $F(53,824) = 27.30$ ,  $p < 0.001$ ) which is not much larger in substantial terms, but because of the large  $n$  it is statistically significant ( $F(26,824) = 3.94$ ,  $p < 0.001$ ). Figure 1 shows the ordinary least square regression lines for each subject. From this figure, we would probably not feel comfortable giving a precise prediction of emotional impact for each photograph from knowing

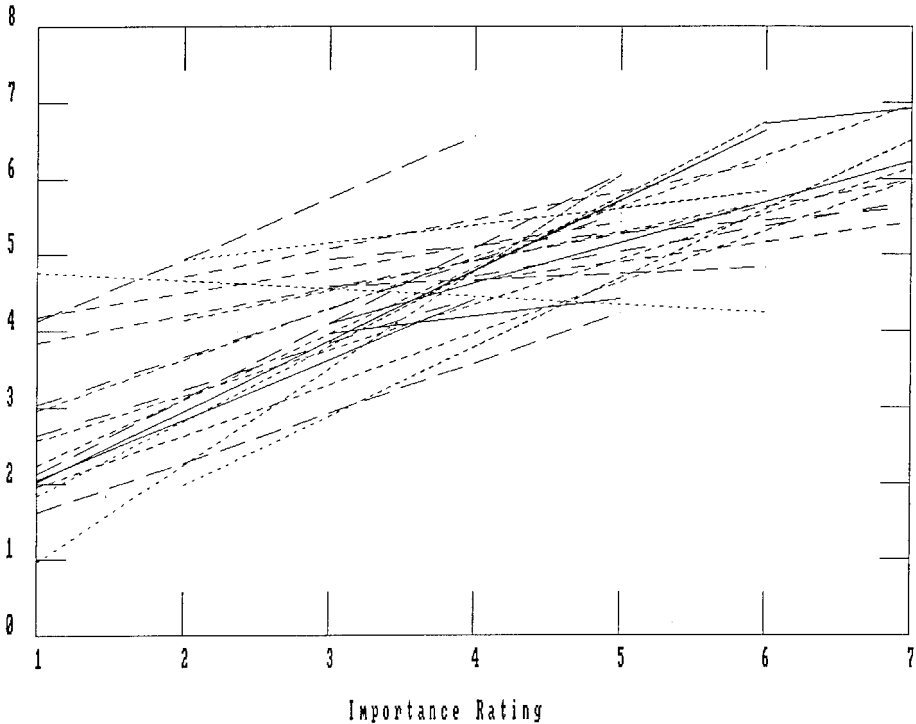


Figure 1. The ordinary least squares regression lines for each person, using importance to predict emotion

the importance rating, even knowing whose picture it is. This figure is particularly scattered because many of the people had a relatively small number of correct responses and therefore some regression lines are unreliable. As with the analyses for reaction times, this approach treats photographs as a random sample from some population.

The final approach also treats photographs as a random sample from each person's summer but also treats subjects as a random sample from some population. In this section I will go through two relatively simple models. In the Appendix I describe some more complex models. In the first model, sometimes called the random intercept model, the intercept is allowed to vary randomly for subjects. It assumes that the  $u_j$  are Normally distributed. The result is

$$emot_{ij} = 4.47 + 0.60 imp_{ij} + u_j + e_{ij}$$

(0.12) (0.03)

with  $\text{var}(u_j) = 0.35$  ( $se = 0.10$ ) and  $\text{var}(e_{ij}) = 1.00$  ( $se = 0.05$ ). The size of  $\text{var}(u_j)$  compared with its standard error and its size relative to  $\text{var}(e_{ij})$  indicate the importance of using a multilevel model as opposed to the first approach where it was assumed that people did not differ. The  $\chi^2$  is 2559.16.

As was done for the model depicted in Figure 1, it was felt that the slopes might also differ for individuals. The slope was allowed to vary randomly for subjects. This

was done by using an additional variance term,  $v_j$ , to be the variation about  $\beta_1$  and assuming the  $v_j$  are Normally distributed:

$$emot_{ij} = (\beta_0 + u_j) + (\beta_1 + v_j)imp_{ij} + e_{ij} = \beta_0 + \beta_1 imp_{ij} + u_j + v_j imp_{ij} + e_{ij}$$

The estimated solution is

$$emot_{ij} = 4.55 + 0.57 imp_{ij} + u_j + v_j imp_{ij} + e_{ij}$$

(0.12) (0.06)

with  $\text{var}(u_j) = 0.30$  (se = 0.10),  $\text{var}(v_j) = 0.07$  (se = 0.03),  $\text{cov}(u_j, v_j) = -0.05$  (se = 0.04) and  $\text{var}(e_{ij}) = 0.92$  (se = 0.05). The fit of this model is  $\chi^2 = 2518.65$  which is a significant improvement over the last ( $\chi^2(2) = 40.51$ ,  $p < 0.001$ ). Comparing the variance estimates with their standard errors this difference appears to be due to the variance of  $v_j$ . The total variance for subjects is

$$\begin{aligned} \text{var}(u_j + v_j imp_{ij}) &= \text{var}(u_j) + 2\text{cov}(u_j, v_j) imp_{ij} + \text{var}(v_j) imp_{ij}^2 \\ &= 0.30 + 2(-0.05) imp_{ij} + 0.07(imp_{ij})^2 \end{aligned}$$

Therefore, since  $\text{var}(v_j)$  is with the quadratic term of  $imp_{ij}$ , the relationship between importance and subject variance is quadratic. The variation among subjects is largest when importance scores are either high or low. In the Appendix this model is further evaluated as an illustration of the types of questions that can be posed with multilevel modelling.

While it is simple to quote a single correlation between two variables, and to conclude that they are positively related, with data as rich as collected by Burt *et al.* (1995), further modelling can be used to understand the relationship more fully. The second approach, aggregating the data, confounds several possible effects and is not asking the appropriate question if the research is interested in the relation between emotive and important memories. The third and fourth approaches each allow for differences among people. Each assumes that the photographs are a random sample from people, but the fourth approach also assumes subjects are a random sample of some population. This is a common assumption in all areas of psychology, and allows generalization to other people. The fourth approach also does not require as many parameters to be estimated and therefore is more parsimonious. Further, it is more easily extended as is illustrated in the Appendix.

### Predicting remembering

Many researchers (e.g. Anderson and Conway, 1993; Conway *et al.*, 1996; Reiser *et al.*, 1985; Wagenaar, 1986) have tried to determine the best cues for retrieving memories. In Burt *et al.* the photographs were classified by whether they depicted a participant, a location and/or an activity. Here I use these three as dummy variables to predict whether the person remembers the event. Using these three dummy variables is a different approach from that used by Burt *et al.* who used a seven-value categorical variable (for the seven possible combinations of these – no photograph had none of these). Interaction terms can be used to make these approaches equivalent, but here I try first to model remembering with these as main effects, and then test for

interactions. The hope is for a more parsimonious model which would imply the effects of having a participant, an activity and a location are additive with the logit of remembering.

The most common way to model binary data of this type is to run a logistic regression (see Menard, 1995; Wright, 1997b). Here it would have the form

$$\text{logit } \pi_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 \text{part}_i + \beta_2 \text{act}_i + \beta_3 \text{loc}_i$$

where  $\text{part}_i$ ,  $\text{act}_i$  and  $\text{loc}_i$  are dummy variables for whether a participant, an activity and/or a location are depicted in the photograph and  $\pi_i$  is the probability of remembering. Approximately 95% of the photographs had a location so this is not a very discriminating variable, but will be included to allow comparisons with the original analyses. The actual responses, call them  $\text{mem}_i$ , are  $\pi_i + e_i$ , where the  $e_i$  are usually assumed to be binomially distributed. When solving this using the first approach we get (standard errors below in parentheses):

$$\text{logit } \pi_i = -0.44 + 0.31 \text{part}_i + 0.77 \text{act}_i + 0.55 \text{loc}_i$$

(0.27) (0.13)            (0.12)        (0.26)

which shows that having an activity significantly increases the probability of remembering the event. From SPSS, the test statistic to measure whether this model fits above chance is  $\chi^2(3) = 64.54$  ( $p < 0.001$ ). The amount left over for the residual  $\chi^2$  is 1664.76 so the proportion of  $\chi^2$  deviation accounted for is statistically significant but not substantial. Further models with interactions were estimated. Adding all the second-order interactions did not improve the fit significantly  $\chi^2(3) = 4.98$  ( $p = 0.17$ ), so I will concentrate on the main effects model.

For the second approach, the data were aggregated in two ways. Following the logic of the analyses on reaction times, it is possible to calculate the proportion that each person correctly identified the photograph as their own for each of the seven combinations of activity, location and participant (all photographs had at least one of these). However, because of missing values, listwise deletion would leave only three cases for analyses. The alternative used was to compare the proportion correct for photographs with activities (mean = 0.39, se = 0.03), locations (mean = 0.64, se = 0.04) and participants (mean = 0.50, se = 0.04) ( $F(2,52) = 58.19$ ,  $p < 0.001$ ). This is not entirely justified because a number of photographs had more than one of these, and therefore are counted twice.

Another model that can be tested with the aggregate data involves calculating the proportion of photographs for each person that depict people, activities and locations, and seeing how these predict the overall proportion of remembering. As with the reaction time analyses, this is not a direct way of examining the hypotheses of interest. It yields  $R^2 = 0.12$  ( $F(3,23) = 1.08$ ,  $p = 0.38$ ).

The third approach involves the 26 subject dummy variables. Including just these variables improves the fit significantly over just having a single constant ( $\chi^2(26) = 213.22$ ,  $p < 0.001$ ). The residual is  $\chi^2 = 1516.08$ . Adding the three main effect variables for picture contents increases the fit by an additional  $\chi^2(3) = 60.43$ , which is significant with  $p < 0.001$ . The parameter estimates are similar to those found with the first

approach: 0.89 (se = 0.14) for activity, 0.46 (se = 0.29) for location and 0.29 (se = 0.15) for participants. Again, only activity has a substantial effect.

The final approach is a multilevel logistic regression. There has been considerable discussion about the estimation of multilevel logistic regressions (Breslow and Clayton, 1993; Goldstein, 1991; Paterson, 1995; Rodríguez and Goldman, 1995). Here, second order Taylor expansion is used with predictive quasi-likelihood. Together, these produce good estimates of the fixed parameters (Goldstein, 1995) and with the fairly high number of photographs per person, should produce good estimates for the random parameters also (Wright, 1997a).

The logistic model estimated was

$$\text{logit}\pi_{ij} = -0.24 + 0.30\text{part}_{ij} + 0.88\text{act}_{ij} + 0.47\text{loc}_{ij}$$

(0.34) (0.15)            (0.14)            (0.28)

with  $\text{var}(u_j) = 0.84$  (se = 0.26). The  $\chi^2$  associated with this is 1484.57. In this model, it is assumed that the photograph level residuals follow a binomial distribution. Extra-binomial variation is important to check since it can be a sign of model misspecification (see Wright, 1997a, for detailed discussion). When this assumption was relaxed,  $\chi^2$  dropped to 1482.63, which is not significantly different from the previous model. Thus, the assumption of binomial variation in the residuals of photographs cannot be rejected. This is a good sign for the validity of the model. In a similar way as was done for the reaction times and importance/emotion relationships, further models can be estimated here. However, this relatively simple model provided an adequate fit, and was not greatly improved with more complicated models.

## DISCUSSION

In much psychology research the data that are collected have a multilevel structure. This is particularly true for researchers interested in autobiographical memory. Often researchers have ignored this structure, and thus their data do not satisfy the assumptions of their statistical tests. The purpose of this paper was to explore several possible approaches for analysing data of this type. With this first approach it is assumed that the photographs were randomly sampled from some population and that each is independent of all the others. Because there were substantial differences among people, the independence assumption was not valid for any of the three sets of analyses. Of more importance, researchers may fail to think about the hierarchical structure of their data. For Burt *et al.*'s (1995) data set this means that attention may not have been drawn to individual differences.

The aggregating approach is attractive to many researchers because it does not have the problem of the tendency for the reported *p*-values being too low. Psychologists tend to be very concerned with Type I errors so often adopt conservative methods even when they do not match their research questions. This, however, can greatly increase Type II errors. Cohen (1992) has argued convincingly that researchers should put more emphasis on Type II errors (and effect size) than is currently done. If the choice of using aggregate approaches was simply a matter of power, then arguments could be made for their use. The real problem is that the effects which are estimated

are about the aggregated cases and should not be used to inform beliefs about the lower level units. The examples used here demonstrate the problems trying to decide what relationships the effects are actually measuring. Robinson's (1950) advice to avoid the ecological fallacy should be heeded. It is, of course, worth mentioning that for some research questions we would be interested in differences among people. For these, the second approach would be preferable to the first. However, both the third and fourth approaches allow both these levels to be explored.

The third approach, treating the higher-level cases as  $J-1$  dummy variables, has some uses. Its problems include that estimating the additional parameters makes the model less parsimonious. The large number of parameters puts restrictions on the complexity of the models that can be examined. These concerns are greatest when there are many higher-level cases. If there are only a few higher-level cases, then there is less concern. This approach is conceptually the same as the traditional way in which analyses of covariance are conducted, but using the dummy variables as covariates. It is important to note that this approach treats the photographs as a random sample from some population, but not the subjects. This means researchers should take care in generalizing their results to other people.

The critical step from the third (fixed effect) to the fourth (multilevel) approach involves assumptions made about the distribution of higher-level residuals. This allows fewer parameters to be estimated (the number depending on the assumed distribution) and is more useful for predictive purposes (providing the distributional assumptions are warranted). Another advantage of the multilevel approach is that it forces the analyst to examine additional variance terms, encouraging more thorough exploration thereby allowing more complete descriptions of the data. Methodologically, subjects are treated as a random sample from a population. This is true for most statistical tests, but is a difference between the third and fourth approaches.

It can be argued that if the higher-level cases are not randomly sampled from some population then the fixed approach should be used. This is similar to Cohen's (1976) concern about Clark's (1973) encouragement of using random effects ANOVAs, but his point was whether the choice of stimuli was random. Randomness is critical in statistics as it allows inference from a sample to a population. This is true at each level of the data and in studies with only level. While in an ideal world psychologists would use random sampling (or some form of probability sampling) from well-defined populations, this seldom happens.

With respect to lower-level units, all the approaches assume that the photographs are a random sample from the population of interest. There is no simple alternative to this, so the question becomes defining the population of interest. For Brewer's (1988) study, which had a beeper go off at random times during the day to determine the events, the population for each individual could conceivably be all events that he or she experienced, but for most autobiographical memory studies the population is less clear. The approach I have taken here is to consider the sample as a random sample of some, most likely unknown, population. For Burt *et al.*'s (1995) data the population is the events that people are likely to take non-blurry photographs of during a summer holiday. Of course, the same logic is implicitly taken when subjects are treated as a random sample from a population in most psychology research.

There is now a large literature, especially in education and sociology, demonstrating some of the inadequacies of traditional methods and arguing for multilevel methods. The acceptance of multilevel methods is rapidly growing in these disciplines

particularly because the nested structure is clear; we can see that the pupils are within the classrooms. It is perhaps less intuitive that memories are nested within a person. Further, the memory system has several levels in the hierarchy (e.g. Barsalou, 1988; Conway and Bekerian, 1987; Schooler and Herrmann, 1992) reflecting the nesting of events themselves (Neisser, 1986). There will also be complex relationships among memories at the same level and between different levels. Theoretically these can be incorporated into the statistical models. However, the current theories about the organization of autobiographical memory are not detailed enough to justify building in these factors.

No statistical technique is optimal for all situations. Statistical methods evolve and must be compared with existing ones to identify when each should be used. In this paper multilevel modelling was shown to have several advantages over treating the data as independent, aggregating the data and, to a lesser extent, over treating the subjects as fixed effects.

### ACKNOWLEDGEMENTS

This paper was prepared with support from a British Academy Fellowship. Much thanks to Chris Burt for sending me his data and for discussion about the topic. Thanks also to Martin Conway and two reviewers for helpful comments.

### REFERENCES

- Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, **105**, 290–301.
- Anderson, S. J. and Conway, M. A. (1993). Investigating the structure of autobiographical memories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 1178–1196.
- Barsalou, L. W. (1988). The content and organization of autobiographical memories. In U. Neisser and E. Winograd (Eds) *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp.193–243). New York: Cambridge University Press.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.
- Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser and E. Winograd (Eds), *Remembering reconsidered: Ecological and traditional approaches to memory* (pp.21–90). Cambridge: Cambridge University Press.
- Bryk, A. S. and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, **101**, 147–158.
- Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications.
- Burt, C. D. B. (1992). Retrieval characteristics of autobiographical memories: Event and data information. *Applied Cognitive Psychology*, **6**, 389–404.
- Burt, C. D. B., Mitchell, D. A., Raggatt, P. T. F., Jones, C. A. and Cowan, T. M. (1995). A snapshot of autobiographical memory retrieval characteristics. *Applied Cognitive Psychology*, **9**, 61–74.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335–359.
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, **114**, 494–509.

- Cliff, N. (1994). Predicting ordinal relations. *British Journal of Mathematical and Statistical Psychology*, **47**, 127–150.
- Clogg, C. C. and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. London: Sage Publications.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, **70**, 426–443.
- Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior*, **15**, 261–262.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155–159.
- Conway, M. A. (1993). Method and meaning in memory research. In G. M. Davies and R. H. Logie (Eds), *Memory in everyday life* (pp. 499–524). Amsterdam: North-Holland.
- Conway, M. A. and Bekerian, D. A. (1987). Organization in autobiographical memory. *Memory & Cognition*, **15**, 119–132.
- Conway, M. A., Collins, A. F., Gathercole, S. E. and Anderson, S. J. (1996). Recollections of true and false autobiographical memories. *Journal of Experimental Psychology: General*, **125**, 69–95.
- Crovitz, H. F. and Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society*, **4**, 517–518.
- Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, **98**, 254–267.
- Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, **76**, 622–623.
- Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika*, **78**, 45–51.
- Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.
- Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society: A*, **157**, 317–356.
- Hox, J. J. (1995). *Applied multilevel analysis* (2nd edition). Amsterdam: TT-Publikaties.
- Kreft, I. G. G. and de Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications.
- Kreft, I. G. G., de Leeuw, J. and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, **30**, 1–21.
- Kreft, I. G. G., de Leeuw, J. and van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, and VARCL. *The American Statistician*, **48**, 324–335.
- Larsen, S. F. (1992). Potential flashbulbs: Memories of ordinary news as the baseline. In E. Winograd and U. Neisser (Eds) *Affect and accuracy in recall: Studies of 'flashbulb' memories* (pp. 32–64). Cambridge: Cambridge University Press.
- Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.
- Longford, N. T. (1993). *Random coefficient models*. Oxford: Oxford University Press.
- Menard, S. (1995). *Applied logistic regression analysis* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no.07-106). Thousand Oaks, CA: Sage.
- Neisser, U. (1986). Nested structure of autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–81). Cambridge: Cambridge University Press.
- Paterson, L. (1995). Entry to university by school leavers. In G. Woodhouse (Ed.), *A guide to MLn for new users* (pp. 87–101). Multilevel Models Project. Institute of Education, University of London.
- Perfect, T. J. (1994). What can Brinley plots tell us about cognitive aging? *Journals of Gerontology*, **49**, 60–64.
- Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, **48**, 359–370.
- Reiser, B. J., Black, J. B. and Kalamarides, P. (1985). Knowledge structures in the organization and retrieval of autobiographical memories. *Cognitive Psychology*, **17**, 89–137.

- Robinson, J. A. (1976). Sampling autobiographical memory. *Cognitive Psychology*, **8**, 578–579.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociology Review*, **31**, 106–128.
- Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: A*, **158**, 73–89.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. C. and Kozin, M. (1984). Vivid memories. *Cognition*, **16**, 81–95.
- Saris, W. E. (1988). *Variation in response functions: A source of measurement error in attitude research*. Amsterdam: Sociometric Research Foundation.
- Scariano, S. and Davenport, J. (1987). The effects of violations of the independence assumptions in the one way ANOVA. *The American Statistician*, **41**, 123–129.
- Schooler, J. W. and Herrmann, D. J. (1992). There is more to episodic memory than just episodes. In M. A. Conway, D. C. Rubin, H. Spinnler and W. A. Wagenaar (Eds), *Theoretical perspectives on autobiographical memory* (pp.241–261). London: Kluwer Academic Press.
- Thompson, C. P. (1982). Memory for unique personal events: The roommate study. *Memory & Cognition*, **10**, 324–333.
- Thompson, C. P., Skowronski, J. J. and Betz, A. L. (1993). The use of partial temporal information in dating personal events. *Memory & Cognition*, **21**, 352–360.
- Van der Geer, J. P. (1993a). *Multivariate analysis of categorical data: Applications*. London: Sage Publications.
- Van der Geer, J. P. (1993b). *Multivariate analysis of categorical data: Theory*. London: Sage Publications.
- Vancouver, J. B., Millsap, R. E. and Peters, P. A. (1994). Multilevel analysis of organizational goal congruence. *Journal of Applied Psychology*, **79**, 666–679.
- Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, **18**, 225–252.
- Woodhouse, G. (Ed.) (1995). *A guide to MLn for new users*. Multilevel Models Project. Institute of Education, University of London.
- Woodhouse, G., Rasbash, J., Goldstein, H. and Yang, M. (1995). Introduction to multilevel modelling. In G. Woodhouse (Ed.), *A guide to MLn for new users* (pp.9–57). Multilevel Models Project. Institute of Education, University of London.
- Wright, D. B. (1997a). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Mathematical and Statistical Psychology*, **50**, 21–29.
- Wright, D. B. (1997b). *Understanding statistics: Introduction to statistics for the social sciences*. London: Sage Publications.
- Wright, D. B., Gaskell, G. D. and O’Muircheartaigh, C. A. (1994). How much is ‘Quite a bit’? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, **8**, 479–496.
- Wright, D. B., Gaskell, G. D. and O’Muircheartaigh, C. A. (1998). Flashbulb memory assumptions: Using National surveys to explore cognitive phenomena. *British Journal of Psychology*, **89**, 103–121.
- Wright, D. B. and McDaid, A. T. (1996). Comparing system and estimator variables using data from real line-ups. *Applied Cognitive Psychology*, **10**, 75–84.

## APPENDIX

When exploring the relation between importance and emotion, the final model discussed in the text was the random slope model. Here, this relationship is explored further in order to illustrate the utility and flexibility of the approach. The residuals of this model for photographs are graphed with importance in Figure 2. They are more spread out at lower values. This is a sign of heteroscedascity at the level of photographs. To model this, a term for heteroscedascity was added into the model by

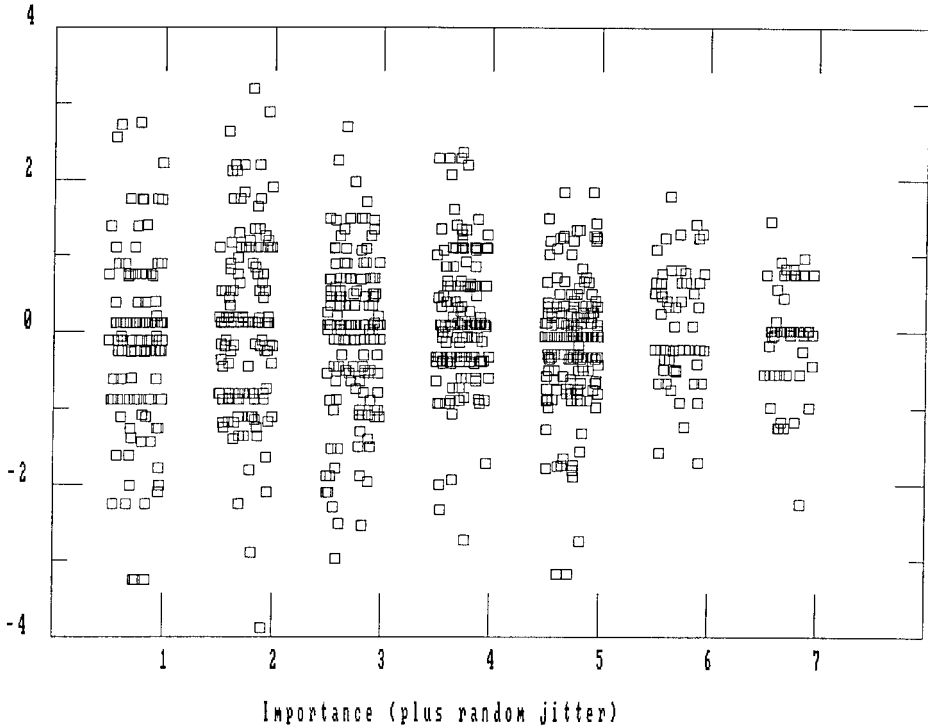


Figure 2. The residuals of emotion ratings for each level of importance. A random 'jitter' has been added to the values of importance so that all the points can be seen

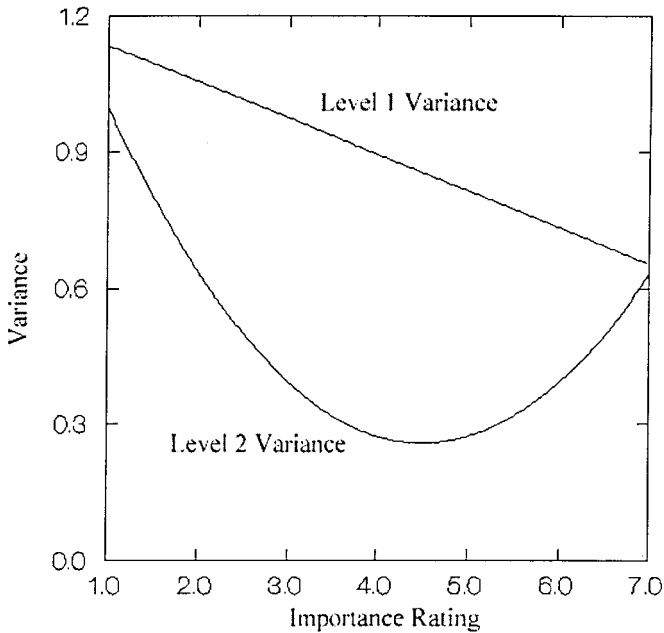


Figure 3. The photograph-level and subject-level residual variances as functions of importance

making a second variance term for photograph residuals, call it  $e2_{ij}$ , and multiplying it by  $imp_{ij}$ .

$$emot_{ij} = \beta_0 + \beta_1 imp_{ij} + u_j + v_j imp_{ij} + e1_{ij} + e2_{ij} imp_{ij}$$

From Figure 2, the variance appears to be a linear function of importance (statistical tests support this). Therefore,  $var(e2_{ij})$ , the quadratic term for calculating total variance in photograph residuals, is constrained to zero. The estimated solution is

$$emot_{ij} = 4.55 + 0.57 imp_{ij} + u_j + v_j imp_{ij} + e1_{ij} + e2_{ij} imp_{ij}$$

(0.12) (0.06)

with  $var(u_j) = 0.30$  (se = 0.10),  $var(v_j) = 0.06$  (se = 0.02),  $cov(u_j, v_j) = -0.05$  (se = 0.03),  $var(e1_{ij}) = 0.92$  (se = 0.05) and  $cov(e1_{ij}, e2_{ij}) = -0.08$  (se = 0.01). This model has the best fit of any of those examined ( $\chi^2 = 2472.16$ , change in  $\chi^2(1) = 46.49$ ,  $p < 0.001$ ). The total variances for both subjects and photographs are shown in Figure 3. The variation for subjects is high for people who gave low or gave high ratings for importance. The variation for photographs decreases as importance goes up. The substantive interpretation of this would be that as an event's importance increases it becomes a better predictor of emotional level. The interpretation for the subject level variance is that the prediction of overall emotion for the summer holiday is poor for people who had, overall, either unimportant or very important holidays.